Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2017

Crossing the Border Twice: Reimporting Prepositions to Alleviate L1-Specific Transfer Errors

Graën, Johannes ; Schneider, Gerold

Abstract: We present a data-driven approach which exploits word alignment in a large parallel corpus with the objective of identifying those verb- and adjective-preposition combinations which are difficult for L2 language learners. This allows us, on the one hand, to provide language-specific ranked lists in order to help learners to focus on particularly challenging combinations given their native language (L1). On the other hand, we provide extensive statistics on such combinations with the objective of facilitating automatic error correction for preposition use in learner texts. We evaluate these lists, first manually, and secondly automatically by applying our statistics to an error-correction task.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-137099
Conference or Workshop Item
Published Version

The following work is licensed under a Creative Commons: Attribution 4.0 International (CC BY 4.0) License.

Originally published at:
Graën, Johannes; Schneider, Gerold (2017). Crossing the Border Twice: Reimporting Prepositions to Alleviate L1-Specific Transfer Errors. In: Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition, Gothenburg, 22 May 2017. Linköpings universitet Electronic Press, 18-26.

Crossing the Border Twice: Reimporting Prepositions to Alleviate L1-Specific Transfer Errors

Johannes Graën
Institute of Computational Linguistics
University of Zurich
graen@cl.uzh.ch

Gerold Schneider
Institute of Computational Linguistics
University of Zurich
gschneid@cl.uzh.ch

Abstract

We present a data-driven approach which exploits word alignment in a large parallel corpus with the objective of identifying those verb- and adjective-preposition combinations which are difficult for L2 language learners. This allows us, on the one hand, to provide language-specific ranked lists in order to help learners to focus on particularly challenging combinations given their native language (L1). On the other hand, we provide extensive statistics on such combinations with the objective of facilitating automatic error correction for preposition use in learner texts. We evaluate these lists, first manually, and secondly automatically by applying our statistics to an error-correction task.

1 Introduction

Computational Linguistics and Learner Error research have made impressive progress recently, but they have not reached their collaborative potential yet (Granger and Lefer 2016, p. 281). For example, while language teaching materials contain lists of idioms and phrasal verbs, the decision for which items to include often does not take actual frequency of use or particular difficulties for learners with specific backgrounds into account.

The current paper addresses this shortcoming, by exploiting large parallel and error-annotated learner corpora. We focus on verb-preposition combinations (VPC), including phrasal verbs and adjective-preposition combinations (APC) obtained from a large parallel corpus (Europarl). For brevity's sake we only describe VPC here.

Our aim is to provide practical and customized help to the learner of a language, here English, by pointing out errors that are likely to be made and to correct them where they occur. In particular, we provide a) a list of VPC/APC that vary considerably between languages, b) a list of specific VPC/APC errors that are to be expected from a native speaker of a particular language, and c) a resource which detects probably incorrect VPC/APC uses and suggests a correction. Concerning c), advances have been made recently due to the CoNLL shared tasks on grammatical error correction (Ng, M. S. Wu, Y. Wu, et al. 2013; Ng, M. S. Wu, Briscoe, et al. 2014), and due to systems targeting preposition errors (Tetreault and Chodorow 2008; Boyd et al. 2012). We evaluate our results on ICLE (Granger, Dagneaux, et al. 2002), the FCE dataset (Yannakoudakis et al. 2011), and the NICT Japanese Learner English Corpus[1]. Furthermore, we exploit ICLE in combination with the British National Corpus (BNC) (Aston and Burnard 1998) to attain collocation statistics which allow us to evaluate the proposed suggestions for corrections.

Non-standard uses by language learners, which we refer to as errors here, can be found at any linguistic level. Some errors can be detected easily by current word-processing tools (e.g. spelling errors) or by re-reading, or consulting dictionaries. But particularly in areas where grammar and lexis interact, there is typically a lack of tools.

One frequent source of lexico-grammatical errors are VPC. While semantically transparent prepositions (e.g. *stand on*) are relatively stable cross-linguistically, the frequent nonsemantic prepositions (e.g. *wait for*) and phrasal verbs (e.g. *depend on*) show enormous cross-linguistic variation. VPC are difficult to acquire for language learners (Gilquin, Granger, et al. 2011, pp. 59–60). Phrasal verbs represent "one of the most notoriously challenging aspects of English language instruction" (Gardner and Davies 2007, p. 339; see

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

[1] https://alaginrc.nict.go.jp/nict_jle/index_E.html

Johannes Graën and Gerold Schneider 2017. Crossing the border twice: Reimporting prepositions to alleviate L1-specific transfer errors. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017.* Linköping Electronic Conference Proceedings 134: 18–26.

also Gilquin 2015). In the CoNLL shared tasks, prepositional errors were the third most frequent error type at 5 to 9 % of all errors (only determiner errors and noun number are more frequent). We include APC as they are often similarly difficult to acquire for learners of English. Benson et al. (2009) recognize APC as an independent category in addition to VPC.

2 Corpus Preparation

We extracted parallel text units in English, Finnish, French, German, Italian, Polish and Spanish from the *Corrected & Structured Europarl Corpus (CoStEP)* (Graën et al. 2014) which is a cleaned version of the Europarl Corpus (Koehn 2005).

We identified approximately 40 million tokens in five languages: English, French, German, Italian and Spanish. Finnish and Polish have considerable fewer tokens than the other languages (30 million and 10 million, respectively).[2]

2.1 Tagging and Lemmatization

For tagging and lemmatization, we used *TreeTagger* (Schmid 1994). To increase tagging accuracy for words unknown to the language model, we had to extend the tagging lexica, especially the German one, with lemmas and part-of-speech tags for frequent words. Moreover, we used the word alignment information between the languages (see below) to disambiguate lemmas for those tokens where the TreeTagger provided multiple lemmatization options.[3]

2.2 Alignment

On the sentence segments identified (about 1.7 million per language), we performed pairwise sentence alignment with *hunalign* (Varga et al. 2005) and based on that word alignment with *GIZA++* (Och and Ney 2003; Gao and Vogel 2008) and the Berkeley Aligner (Liang et al. 2006). While the Berkeley Aligner computes bidirectional word alignments, the alignments of GIZA++ are unidirectional and thus need to be symmetrized if bidirectional alignments are required. We chose the union symmetrization method since it increases recall. Word alignment was performed on the types of all tokens and on lemmas of content words.[4] For the latter, we mapped the individual tag sets to the universal tagset defined by Petrov et al. (2012) and defined content words to be those tokens being tagged as nouns, verbs, adjectives or adverbs.

2.3 Parsing

We used *MaltParser* (Nivre et al. 2006) to derive syntactic dependency relations in English. For parsing our tagged texts, we had to map several part-of-speech tags beforehand as the standard English parameter file distributed with *TreeTagger* slightly differs.[5]

3 Methods

In the following, we first present our concept of backtranslating prepositions based on automatic annotation and alignment frequencies. We then apply it to VPC and introduce our method for error correction.

3.1 Distributions

In a first step, we calculate a lemma distribution matrix by aggregating lemma counts on token alignments. This matrix tells us the translation ratio of each lemma. Each cell contains the probability of a lemma in the source language to be translated into a lemma in the foreign language. For example, the English verb *suffer* is translated to German *leiden* in 42 % of the cases.

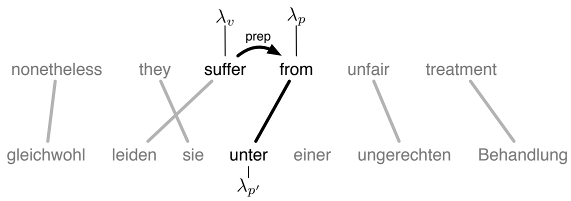

Figure 1: Corpus example: first, the VPC is identified employing syntactic dependency relations, second, the foreign language preposition of the VPC is retrieved following the word alignment.

We then retrieve the set of all English VPC (consisting of verb λ_v and preposition λ_p with the verb showing a syntactic 'preposition' relation to the preposition as depicted in Fig. 1) and calculate the

[2]For Polish this is due to the fact that Poland joined the European Union in 2004 and translations for the debates are only available from 2006 onwards. In case of Finnish, the gap can be explained by the language itself which features a rich morphology, thus resulting in less tokens at the expense of more word forms.

[3]The disambiguation approach is similar to the one Volk et al. (2016) describe, except that we combine alignment information for all languages simultaneously.

[4]If no lemma provided we used the word form instead.

[5]It distinguishes e.g. between main and auxiliary verbs, and between prepositions and complementizers.

distribution of observed prepositions. For example, the English verb *suffer* occurs with the preposition *from* in 26 % of all cases, but also, more rarely, with other prepositions.[6] We do not attempt to make a distinction between phrasal verbs, PP complements or PP adjuncts in our data-driven approach.

For each VPC, we count the foreign prepositions $\lambda_{p'}$ as they are aligned with the source VPC's prepositions λ_p.[7] We do this step for each language separately.

3.2 Backtranslation Score (BTS)

By multiplying these foreign prepositions with the lemma distribution matrix, we obtain a list of English prepositions and values that we call backtranslation score (BTS). BTS tells us how preferred a certain source language preposition is[8] for a foreign language, given a particular VPC.

3.3 Backtranslation Ratio (BTR)

We then normalize BTS to what we refer to as backtranslation ratio (BTR), such that the BTR of the correct English preposition for a particular verb and language is 1.0, i.e. each preposition's BTS divided by the BTS of the correct original preposition, which is shown in Table 1. A BTR above 1.0 indicates that it is more likely to choose a wrong preposition than the correct one, according to our language model, which is based on alignment (see Appendix for the most likely incorrect preposition per language, with their BTR).

The BTR calculated for English VPC give us an impression of how difficult the preposition of a particular expression would be for a speaker of the respective language. For instance, the highest BTR for the verb *aim* is 2.74 for German (preposition *on*, presumably due to German *zielen auf*) and 2.81 for French (preposition *in*, indirectly due to French *viser* + object, see next subsection) while *at* is 1.0 by definition.

We also include the raw frequency of VPC and derive the final ranking for each VPC and language based on both normalized scores.[9] For space reasons, we only present the intersection of all language specific VPC and APC lists in Table 2.

λ_v	λ_p	$\lambda_{p''}$	BTS	BTR
suffer	from	under	102.512	2.51
suffer	from	of	100.036	2.46
suffer	from	in	78.559	1.93
suffer	from	by	51.188	1.25
suffer	from	on	46.534	1.14
suffer	from	**from**	**40.966**	**1.00**
suffer	from	with	36.322	0.89
suffer	from	among	27.927	0.68
suffer	from	at	15.791	0.39
suffer	from	amongst	11.207	0.28
		⋮		

Table 1: Backtranslation score (BTS) and backtranslation ratio (BTR) for different backtranslated prepositions ($\lambda_{p''}$) of *suffer from*.

3.4 Suggestions for Corrections

In addition to lists of difficult VPC and APC, we also suggest a correction for incorrect combinations based on the distribution of prepositions retrieved. Errors can be simple misproductions such as typos or copy-paste errors, which are typically spotted when carefully re-reading a text. But when speakers of certain linguistic backgrounds keep producing the same non-standard form repeatedly, often due to native language influence such as transfer, they make errors which are more difficult to detect for them, and thus a resource which spots these is particularly helpful. These errors follow a repeated pattern, often reaching collocational status. Schneider and Gilquin (2016) use collocation-based statistics to detect such non-standard VPC by measuring the expected (E) collocational strength in Learner English (based on the International Corpus of Learner English (ICLE)), compared to the observed (O) collocational strength in native English (based on the BNC).

$$\text{O/E-ratio} = \frac{\text{O/E(ICLE)}}{\text{O/E(BNC)}} \tag{1}$$

We detect VPC errors following the same method, then address the question if we can provide the appropriate correction. Given an incorrect VPC, we suggest the most likely preposition, given the verb. As some errors involve a preposition instead of a direct object, our algorithm suggests to

[6] The second most frequent preposition together with *suffer* is *in*, occurring in 9 %. In 2 %, *suffer* is modified by a PP headed by *under*.

[7] We only consider alignments from English prepositions to prepositions in other languages.

[8] As we multiply by the entire lemma distribution matrix, this could theoretically also be other words than prepositions, but in practice only the prepositions count here.

[9] We calculate the same measures for APC analogously.

VERB/ADJ	PREP	OK?	I	N	F
aim	at	yes	+		
arrive	at	yes	+	+	+
benefit	from	yes	+		
breathe	into	?		*n/a*	
channel	into	yes		*n/a*	
complain	about	yes	+	+	+
compliment	on	yes			
convert	into	yes		*n/a*	
depend	on	yes	+		+
direct	at	yes	+		
divide	into	?		*n/a*	
emanate	from	yes			
embark	on	yes			
enter	into	?		*n/a*	
estimate	at	yes	+		
exclude	from	yes	+		
exempt	from	yes	+		
fall	within	yes			
force	into	yes		*n/a*	
gain	from	yes	+		
hang	over	no		*n/a*	
incorporate	into	?		*n/a*	
integrate	into	?		*n/a*	
level	at	no		*n/a*	
look	at	yes	+	+	+
miss	from	yes			
plunge	into	?		*n/a*	
preside	over	yes			
profit	from	yes	+		
protect	from	yes			
recover	from	yes			
suffer	from	yes			+
talk	about	yes	+	+	+
target	at	yes	+		
throw	into	?		*n/a*	
transform	into	?		*n/a*	
translate	into	?		*n/a*	
transpose	into	?		*n/a*	
wait	for	yes	+	+	+
worry	about	yes			+
absent	from	yes		+	
conditional	on	yes		+	
dependent	on	yes	+	+	+
early	as	no		*n/a*	
exempt	from	yes	+		
sceptical	about	yes	+		
serious	about	yes	+		
Total		34/10/3	23/31		

Table 2: Language-independent VPC/APC obtained by intersecting the language-specific recommendation lists. 23 out of 31 relevant ones can be found in at least one of the learner corpora we searched (I = ICLE; N = NICT; F = FCE).

use a direct object if the raw frequency of a verb is at least twice as high as the number of VPC involving that verb.

4 Results

In the following, we present results for all three aims identified above.

As we cannot give the full lists of recommended language-specific lists here,[10] we will focus instead on three verb-preposition combinations that are particularly useful to concentrate on and to learn for native speakers of German:

- *suffer from*: corresponds to German *leiden unter*, the preposition 'unter' directly translates as 'under'.
- *wait for*: corresponds to German *warten auf*, 'auf' directly translates as 'on'.
- *consist of*: corresponds to German *bestehen aus*, 'aus' directly translates as 'from'.

The recommended lists overlap, yet also differ considerably between languages. The amount of overlapping VPC of the whole lists ranges from 58 % for German-Polish to 97 % for French-Italian, reflecting the typological similarity of the languages. We consider those items that occur in each of the 5 language-specific lists as generally hard to learn. This language-independent list is given in Table 2.

The list of the top true positives, i.e. the correct suggestion for erroneous or non-standard uses of VPC/APC structures from Schneider and Gilquin (2016) is given in Table 3. The first column shows the verb or adjective, the second column the incorrect preposition, the third column the manually corrected preposition. *obj* means that the manual annotation suggests to use a direct object instead of a PP (e.g. *attack against someone* has manually been corrected to *attack someone*), and *n/a* means that the manually suggested correction is more complex, e.g. *diverse by* has manually been corrected to *different according to*. The ultimate column shows whether the automatic correction matches the manual correction.

5 Evaluation

We have evaluated our approach in two ways, which we describe in the following.

[10] We provide the full VPC and APC recommendation lists at `http://pub.cl.uzh.ch/purl/reimporting_prepositions`.

VERB/ADJ	PREP	CORR	MATCH?
accuse	for	of	yes
addict	on	to	yes
alarm	of	at	yes
apply	into	to	yes
assist	to	*obj*	yes
assure	to	*obj*	yes
aspire	for	to	yes
attack	against	*obj*	yes
aware	about	of	yes
belong	into	to	yes
benefit	out	from	yes
call	like	*obj*	no
characterize	with	by	yes
charge	of	with	yes
confront	to	with	yes
consist	on	of	yes
deal	about	with	yes
deprive	from	of	yes
destructive	for	to	yes
discuss	about	*obj*	yes
estimate	to	at	yes
extend	of	to	no
impose	to	on	yes
indulge	into	in	yes
interest	for	in	no
involve	into	in	yes
relate	with	to	yes
replace	to	by	no
resist	to	*obj*	yes
select	among	from	no
separate	between	*n/a*	no
study	about	*obj*	yes
understand	towards	*obj*	yes
view	upon	on	no
bad	to	for	no
capable	in	of	yes
conscious	about	of	yes
critical	against	of	yes
critical	towards	of	yes
dependent	from	on	yes
dependent	of	on	yes
diverse	by	*n/a*	no
guilty	for	of	yes
independent	on	of	yes
responsible	of	for	yes
superior	than	to	yes
synonymous	to	with	yes
worth	for	*obj*	no
Total			38/48

Table 3: Incorrect VPC/APC together with the correction suggested by our algorithm. The list of incorrect VPC/APC structures originates from (Schneider and Gilquin 2016).

First, we have evaluated the list of language-independent suggestions. In column 3 of table 2, we consider an item a true positive if it contains a non-semantic, non-compositional preposition, or if the preposition is language-specific. Precision is at 72 %. Our method does not seem to work reliably on the preposition 'into', which does not exist as a preposition in most languages, but which is semantically transparent. We thus decided to exclude this preposition in the second evaluation, given in columns 4-6, in which we check if errors corresponding to this type occur in learner corpora. 74 % of the remaining combinations are found in at least one of the learner corpora.

Second, we have tested the ability of our method to correct frequent non-standard or erroneous verb- and adjective-preposition combinations. The results are given in Table 3. PREP is the erroneous preposition, CORR the suggested correction by our algorithm, and MATCH? indicates if the suggested correction is correct. The results indicate a precision of 79.2 %, and the upper bound (*n/a* cannot be predicted correctly) is 95.8 %. Some of the errors may stem from the fact that the European parliament uses some fixed phrases that are rare in other registers.

Tetreault and Chodorow (2008) report 80 % precision at 19 % recall on the task of recognizing preposition errors in essays written by non-native students. Boyd et al. (2012) report 40 % F-score on recognizing preposition errors, and 30 % F-score on correcting them. The best performing system on the *prep* type of error in Ng, M. S. Wu, Briscoe, et al. (2014) is Felice et al. (2014), who report about 40 % precision and recall using a combination of a rule-based and an SMT approach. As their task includes both recognizing and correcting preposition errors, and all the above approaches use token-based evaluation while ours is type-based, a comparison is difficult to make, but our results appear to be competitive.

6 Conclusions

We have employed word alignment in a large parallel corpus to identify potentially difficult VPC/APC. We have compiled language-specific ranked lists in order to help learners to focus on particularly challenging combinations given their native language (L1). We have also combined the language-specific findings into a list of generally difficult combinations. As expected, Romance languages exhibit a larger overlap of combinations than German, and Polish is particularly different.

We evaluate our procedure in two ways. First we have manually assessed the precision of the language-independent list, which obtains 72 % precision. Secondly, we apply our method to an error correction task to predict the intended preposition given frequent erroneous VPC or APC. We achieved a precision of 79.2 %.

For future work, we plan to conduct the same calculations for other languages so that we will be able, for instance, to predict potentially erroneous use of German prepositions by native speakers of other languages.

Acknowledgments

This research was supported by the Swiss National Science Foundation under grant 105215_146781/1 through the project "SPARCLING – Large-scale Annotation and Alignment of Parallel Corpora for the Investigation of Linguistic Variation".

References

Aston, Guy and Lou Burnard (1998). *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh University Press.

Benson, Morton, Evelyn Benson, and Robert Ilson (2009). *The BBI combinatory dictionary of English: Your guide to collocations and grammar*. John Benjamins Publishing.

Boyd, Adriane, Marion Zepf, and Detmar Meurers (2012). "Informing Determiner and Preposition Error Correction with Hierarchical Word Clustering". In: *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*. Montreal, Canada: Association for Computational Linguistics, pp. 208–215.

Felice, Mariano, Zheng Yuan, E. Øistein Andersen, Helen Yannakoudakis, and Ekaterina Kochmar (2014). "Grammatical error correction using hybrid systems and type filtering". In: *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*. Baltimore, Maryland: Association for Computational Linguistics, pp. 15–24.

Gao, Qin and Stephan Vogel (2008). "Parallel implementations of word alignment tool". In: *Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP)*. Association for Computational Linguistics, pp. 49–57.

Gardner, Dee and Mark Davies (2007). "Pointing Out Frequent Phrasal Verbs: A Corpus-Based Analysis". In: *TESOL quarterly* 41.2, pp. 339–359.

Gilquin, Gaëtanelle (2015). "The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach". In: *Corpus Linguistics and Linguistic Theory* 11.1, pp. 51–88.

Gilquin, Gaëtanelle, Sylviane Granger, et al. (2011). "From EFL to ESL: evidence from the International Corpus of Learner English". In: *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, pp. 55–78.

Graën, Johannes, Dolores Batinic, and Martin Volk (2014). "Cleaning the Europarl Corpus for Linguistic Applications". In: *Proceedings of the Conference on Natural Language Processing (KONVENS)*. (Hildesheim). Stiftung Universität Hildesheim, pp. 222–227.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot (2002). *International corpus of learner English*. Presses universitaires de Louvain.

Granger, Sylviane and Marie-Aude Lefer (2016). "From general to learners' bilingual dictionaries: Towards a more effective fulfilment of advanced learners' phraseological needs". In: *International Journal of Lexicography*, pp. 279–295.

Koehn, Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: *Machine Translation Summit*. (Phuket). Vol. 5. Asia-Pacific Association for Machine Translation (AAMT), pp. 79–86.

Liang, Percy, Ben Taskar, and Dan Klein (2006). "Alignment by agreement". In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pp. 104–111.

Ng, Tou Hwee, Mei Siew Wu, Ted Briscoe, Christian Hadiwinoto, Hendy Raymond Susanto, and Christopher Bryant (2014). "The CoNLL-2014 Shared Task on Grammatical Error Correction". In: *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*. Baltimore, Maryland: Association for Computational Linguistics, pp. 1–14.

Ng, Tou Hwee, Mei Siew Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault (2013). “The CoNLL-2013 Shared Task on Grammatical Error Correction”. In: *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL): Shared Task*. Sofia: Association for Computational Linguistics, pp. 1–12.

Nivre, Joakim, Johan Hall, and Jens Nilsson (2006). “Maltparser: A data-driven parser-generator for dependency parsing”. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Vol. 6, pp. 2216–2219.

Och, Franz Josef and Hermann Ney (2003). “A Systematic Comparison of Various Statistical Alignment Models”. In: *Computational linguistics* 29.1, pp. 19–51.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2012). “A Universal Part-of-Speech Tagset”. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

Schmid, Helmut (1994). “Probabilistic part-of-speech tagging using decision trees”. In: *Proceedings of International Conference on New Methods in Natural Language Processing (NeMLaP)*. (Manchester). Vol. 12, pp. 44–49.

Schneider, Gerold and Gaëtanelle Gilquin (2016). “Detecting Innovations in a Parsed Corpus of Learner English”. In: *International Journal of Learner Corpus Research* 2.2.

Tetreault, R. Joel and Martin Chodorow (2008). “The Ups and Downs of Preposition Error Detection in ESL Writing”. In: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*. Manchester: COLING 2008 Organizing Committee, pp. 865–872.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón (2005). “Parallel corpora for medium density languages”. In: *Proceedings of the Recent Advances in Natural Language Processing (RANLP)*. (Borovets), pp. 590–596.

Volk, Martin, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel (2016). “Building a Parallel Corpus on the World’s Oldest Banking Magazine”. In: *Proceedings of the Conference on Natural Language Processing (KONVENS)*. (Bochum).

Yannakoudakis, Helen, Ted Briscoe, and Ben Medlock (2011). “A new dataset and method for automatically grading ESOL texts”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*. Vol. 1, pp. 180–189.

A Most relevant VPC for English learners with L1 being German and French

	German				French			
no	λ_v	λ_p	$\lambda_{p''}$	BTR	λ_v	λ_p	$\lambda_{p''}$	BTR
1	think	of	on	1.09	deal	with	of	2.07
2	impose	on	for	1.36	provide	for	of	1.24
3	hope	for	on	1.07	call	for	of	1.82
4	remind	of	on	1.22	decide	on	of	1.05
5	prevent	from	of	1.83	comply	with	of	1.60
6	consist	of	from	1.38	hope	for	of	1.00
7	postpone	until	by	1.06	ask	for	of	2.08
8	exclude	from	of	1.64	face	with	in	1.65
9	aim	at	on	2.74	push	for	of	1.07
10	talk	about	on	3.34	confront	with	in	1.19
11	look	at	in	3.40	cope	with	in	1.47
12	gain	from	of	1.42	reserve	for	in	1.19
13	deliver	on	in	1.37	inflict	on	in	1.11
14	receive	from	of	2.00	spend	on	for	1.75
15	emanate	from	of	1.19	apologise	for	of	1.26
16	compose	of	from	1.30	qualify	for	of	1.15
17	wait	for	on	2.25	strive	for	of	1.32
18	embark	on	in	1.69	associate	with	in	1.92
19	compliment	on	for	1.49	wait	for	of	1.99
20	benefit	from	of	2.72	aim	at	in	2.81
21	shed	on	in	1.62	last	for	of	1.23
22	suffer	from	under	2.44	expire	on	in	1.25
23	dispense	with	on	1.57	allow	for	of	2.28
24	stop	from	of	1.88	arrange	for	of	1.44
25	warn	against	before	1.82	cater	for	of	1.45
26	protect	from	before	2.42	confer	on	in	1.79
27	test	on	in	1.65	look	at	in	5.08
28	abstain	from	in	2.42	account	for	of	2.50
29	hear	from	of	2.65	arrive	at	in	2.77
30	refrain	from	of	2.44	embark	on	in	2.37
31	inform	of	on	2.92	blame	for	of	2.28
32	profit	from	of	2.14	direct	at	in	2.79
33	free	from	of	2.21	destine	for	in	2.27
34	direct	at	on	2.74	estimate	at	in	2.42
35	spend	on	for	3.76	resume	at	in	2.31
36	target	at	on	2.66	burden	with	of	2.28
37	worry	about	on	3.01	concern	with	of	4.26
38	estimate	at	on	2.49	align	with	on	2.59
39	recover	from	of	2.45	fill	with	of	2.69
40	delight	with	on	2.65	congratulate	on	for	6.68
41	depend	on	of	5.01	depend	on	of	5.77
42	arrive	at	in	4.07	search	for	of	2.98
43	exempt	from	of	3.13	level	at	in	3.04
44	differ	from	of	3.62	please	with	of	4.43
45	level	at	on	2.99	care	for	of	3.59
46	depart	from	of	3.24	dispense	with	of	3.34
47	expect	from	of	3.91	forgive	for	of	4.25
48	complain	about	on	3.55	target	at	on	4.74

B Most relevant VPC for English learners with L1 being Spanish and Polish

	Spanish				Polish			
no	λ_v	λ_p	$\lambda_{p''}$	BTR	λ_v	λ_p	$\lambda_{p''}$	BTR
1	thank	for	by	1.01	talk	about	of	1.40
2	deal	with	of	1.36	vote	in	for	2.16
3	call	for	of	1.16	ask	for	of	1.61
4	ask	for	of	1.15	allow	for	on	1.17
5	impose	on	in	1.10	look	at	on	2.29
6	consist	of	in	1.02	deprive	of	by	1.00
7	pass	on	of	1.08	concern	with	of	1.37
8	build	on	in	1.09	wait	for	on	1.47
9	hope	for	of	1.17	hope	for	on	1.24
10	allow	for	of	1.31	learn	from	with	1.48
11	wait	for	of	1.50	remove	from	with	1.33
12	equip	with	of	1.01	pass	on	in	1.48
13	apologise	for	by	1.04	press	for	on	1.14
14	compensate	for	of	1.19	schedule	for	on	1.10
15	think	of	in	1.85	aim	at	on	2.37
16	concern	with	of	1.66	confer	on	in	1.13
17	argue	for	of	1.15	fight	for	of	1.62
18	aim	at	in	2.45	regard	as	for	2.02
19	base	on	in	3.66	compose	of	with	1.10
20	congratulate	on	by	2.53	discriminate	against	of	1.34
21	deliver	on	in	1.21	decide	on	of	1.93
22	qualify	for	of	1.11	depend	on	from	2.17
23	pick	on	of	1.12	avail	of	with	1.09
24	punish	for	by	1.02	fill	with	by	1.16
25	touch	on	in	1.62	benefit	from	with	2.24
26	arrange	for	of	1.24	worry	about	of	1.50
27	elaborate	on	in	1.22	label	of	in	1.22
28	acquaint	with	of	1.32	escape	from	with	1.20
29	destine	for	of	1.48	congratulate	on	in	3.03
30	inflict	on	in	1.55	withdraw	from	with	1.52
31	focus	on	in	4.01	burden	with	of	1.16
32	place	on	in	3.05	dispose	of	in	1.35
33	confer	on	in	1.89	suffer	from	of	2.04
34	cater	for	of	1.77	exclude	from	with	1.98
35	impact	on	in	1.94	emerge	from	with	2.03
36	direct	at	in	2.37	derive	from	with	1.98
37	account	for	of	2.81	originate	from	with	1.64
38	search	for	of	2.04	gain	from	with	1.83
39	rest	on	in	2.17	arise	from	with	2.26
40	arrive	at	in	3.18	exempt	from	with	1.70
41	resume	at	in	2.16	report	on	of	2.12
42	look	at	in	6.99	protect	from	before	2.22
43	dwell	on	in	2.51	recover	from	with	1.64
44	spend	on	in	3.78	release	from	with	1.69
45	concentrate	on	in	4.41	stem	from	with	2.11
46	insist	on	in	4.31	touch	on	in	2.33
47	rely	on	in	4.00	import	from	with	2.13
48	compliment	on	by	2.70	quote	from	with	1.91